Development of algorithm for sentiment analysis to uncover emotions in the text

Synopsis for Registration as Research Student

For

Doctoral Degree

Swami Ramanand Teerth Marathwada University, Nanded

Nanded-431606, Maharashtra State (MS), India.

Swati Bapurao Tondare

Research Scholar

Under Research Guide

Dr. Archana Harsing Sable

School of Computational Sciences,

Swami Ramanand Teerth Marathwada University, Nanded

Abstract of proposed work plan:

A rapid rise in the field of data mining has occurred in recent years. Analysis of Sentiment is an idea that includes numerous happenings including sentiment extraction, classification of sentiment, classification of subjectivity, a summary of the opinion, or spam detection amongst others. Opinion Mining (OM) or Sentiment Analysis (SA) is a mathematical study of views, emotions, sentiments, assessments, and behaviors of people, including products, organizations, services, issues, individuals, topics, events also their attributes. The advancement in web technology tools has given rights to their users to share their opinions on things that they interact with within their daily life. They can post it online on different platforms available. This massive amount of data is used to find out major curiousness that concerns people most and is very beneficial to many firms. To get benefited from this unstructured, inconsistent data written in natural language, first, it needs to be clean and transform and then we will be able to perform different operations like classification and aspect-based sentiment analysis..Now-a-days, the conventional research models are experimented using the public review datasets. However, this kind of review has not been evaluated keenly by concerning the sentiments. By categorizing the sentiments as positive or negative, it will not provide the original and the concealed information beyond the actual concepts of sentiments. In addition, there are some specific sentences that are quite complicating and accurate classification cannot be performed.

We will develop sentiment analysis algorithms that overcome the current problem like insufficient labeled data and incapability to deal with complex sentences, emotional analysis. We will also work on sarcasm detection, double negative detection, and multipolarity.

To train and test proposed model ,web use biggest open-source database for sentiment analysis created by IMDb.

Keywords- Emotion analysis, sentiment analysis,Application-oriented Analysis, Machine Learning Algorithms, Performance Analysis

1. Introduction:

Data Mining (DM) is the method by which the databases gather valuable information. It is used to manage large quantities of data. For the estimation of future developments and behavior, data mining is used. A vast volume of data is difficult to interpret, so DM is utilized. It is also called KDD (knowledge discovery in a database). To forecast data patterns and behavior, data mining techniques are used and allow organizations to make their decisions [1]. Because of its popularity in numerous fields, data mining is becoming popular. It relates to healthcare, telecommunications, finance, education, business also many others [2]. DM is divided into 2 groups- prescriptive mining & descriptive mining. We describe and summarize the current data in the descriptive mining process [3]. SA is also called "Opinion Mining" or "emotional AI," alluding to the use of NLP (natural language processing) and text mining as well as mathematical linguistic & bio- measurement methodically used for emotional & contextual information to be recognized, extricated, measured, and analyzed. SA typically includes the voice of client material, such as web and web-based social networking surveys and reviews [4]. Opinion Mining (OM) is a powerful area of DM that helps organizations in discovering valuable insights into the needs and preferences of their customers. Opinion Mining is intended to classify the portion that expresses the perspective and that which is transmitted. OM is a process of simulated natural language that enables users to monitor the mood of a given commodity or topic. It requires the creation of a mechanism for gathering and reviewing opinions about the commodity generated in blog

articles, blogs, tweets, or comments. It has been researched extensively in data processing, texts, and network processing. [5].

2. Literature Review and Development in the Project:

International status-

In 2019, Saad and Yang [1] have aimed for giving a complete tweet sentiment analysis on the basis of ordinal regression with machine learning algorithms. The suggested model included pre-processing tweets as first step and with the feature extraction model, an effective feature was generated. The methods such as SVR, RF, Multinomial logistic regression (SoftMax), and DTs were employed for classifying the sentiment analysis. Moreover, twitter dataset was used for experimenting the suggested model. The test results have shown that the suggested model has attained the best accuracy, and alsoDTs were performed well when compared over other methods. In 2018, Fang *et al*. [2] have suggestedmulti-strategy sentiment analysis models using the semantic fuzziness for resolving the issues. The outcomes have demonstrated that the proposed model has attained high efficiency.

In 2019, Afzaal *et al*. [3] have recommended a novel approach of aspect-based sentiment Classification, which recognized the features in a precise manner and attained the best classification accuracy. Moreover, the scheme was developed as a mobile application, which assisted the tourists in identifying the best hotel in the town, and the proposed model was analyzed using the real-world data sets. The results have shown that the presented model was effective in both recognition as well as classification.

In 2019, Feizollah *et al*. [4] have concentrated on tweets related to two halal products such as halal cosmetics and halal tourism. By utilizing Twitter search function, Twitter information was extracted, and a new model was employed for data filtering. Later, with the help of deep learning models, a test was performed for computing and evaluating the tweets. Moreover, for enhancing the accuracy and building prediction methods, RNN, CNN, and LSTM were employed. From the outcomes, it was seemed that the combination of LSTM and CNN attained the best accuracy.

In 2018, Mukhtar *et al*. [5] have performed the sentiment analysis to the Urdu blogs attained from several domain with Supervised Machine learning and Lexicon-based models. In Lexicon-based models, a well-performing Urdu sentiment analyzer and an Urdu Sentiment Lexicons were employed, whereas, in Supervised Machine learning algorithm, DT, KNN, and SVM were employed. The data were combined from the two sources for performing the best sentiment analysis. Based on the tests conducted, the outcomes were shown that the Lexicon-based model was superior to the supervised machine learning algorithm.

In 2020, Kumar *et al*. [6] have presented a hybrid deep learning approach named ConVNet-SVMBoVW that dealt with the real-time data for predicting the fine-grained sentiment. In order to measure the hybrid polarity, an aggregation model was developed. Moreover, SVM was used for training the BoVW to forecast the sentiment of visual content. Finally, it was concluded that the suggested ConvNet-SVMBoVW was outperformed by the conventional models.

In 2018, Abdi *et al*. [7] have proffered a machine learning technique for summarizing the opinions of the users mentioned in reviews. The suggested method merged multiple kinds of features into a unique feature set for modelling accurate classification model. Therefore, a performance investigation was done for four best feature selection models for attaining the best performance and seven classifiers for choosing the relevant feature set and recognized an effective machine learning algorithm. The suggested method was implemented in various datasets. The outcomes have demonstrated that the combination of IG as the feature selection approach and SVM-based classification approach enhanced the performance.

In 2019, Ray and Chakrabarti [8] have introduced a deep learning algorithm for extracting the features from text and the user's sentiment analysis with respect to the feature. In opinionated sentences, a seven layer Deep CNN was employed for tagging the features. In order to enhance the performance of sentiment scoring and feature extraction models, the authors merged the deep learning methods using a set of rule-based models. Finally, it was seen that the suggested method achieved the best accuracy. In 2019, Zhao *et al*. [9] have offered a novel image-text consistency driven multimodal sentiment evaluation model, which explored the correlation among the text and image. Later, a multi-modal adaptive sentiment analysis model was implemented. By using the traditional SentiBank model, the mid-level visual features were extracted and those were employed for representing thevisual theories by integrating the different characteristics like social, textual, and visual features for introducing a machine learning model. The suggested model has attained best performance whencompared over traditional models.
In 2019, Park *et al*. [10] have developed a semi-supervised sentiment-discriminative objective for resolving the issue by documents partial sentiment data. The suggested model not only reflected the partial data, but also secured the local structures obtained from real data. The suggested model was evaluated on real time datasets. The results have shown that the suggested model was performing well.

National status-
In 2019, Vashishtha and Susan [11] have calculated the sentiment related to social media posts by a new set of fuzzy rules consisting of many datasets and lexicons. The developed model combined Word Sense Disambiguation and NLP models with a new unsupervised fuzzy rule-based model for categorizing the comments into negative, neutral, and positive sentiment class. The experiments wereperformed on 3 sentiment lexicons, four existing models, and nine freely available twitter datasets.
The outcomes have shown that the introduced method was attaining the best results.
In 2019, Yousif *et al*. [12] have presented a multi-task learning method on the basis of CNN and
RNN. The structure of the suggested method was helpful for denoting the citation context and feature extraction was done in an automatic way. By considering two freely accessible datasets, the suggested technique was analyzed. The outcomes have shown that the proposed model was improved than conventional models.
In 2020, Hassonah *et al*. [13] have recommended hybrid machine learning algorithm for improving the sentiment analysis, because a classification approach was built on the basis of "Positive, Negative, and Neutral" classes with SVM classifier, at the same time two feature selection methods were merged by the MVO and Relief models. Moreover, Twitter data was employed for evaluating the proposed model. The experimental results indicated that the suggested technique was performing well than conventional techniques.
In 2020, Xu *et al*. [14] have introduced a NB method for multi-domain and large-scale Ecommerce platform product review classification of sentiment. Consequently, the parameter evaluation method was extended in NB for continuous learning fashion. Later, for fine-tuning the
learned distribution on the basis of three types of assumptions, many ways were introduced for
acquiring the best performance. The results have shown that the suggested model has high accuracy in Amazon product and movie review sentiment datasets.

In 2018, Smadi *et al*. [15] have proposed existing models on the basis of supervised machinelearning algorithms for specifying the defects of feature-based sentiment analysis of

Arabic hotel'sreview. Moreover, SVM and Deep RNN were developed and trained with word, lexical,morphological, semantic, and syntactic features. The reference dataset of Arabic hotel's review datasetwas used for evaluating the proposed model. The outcomes have shown that SVM was performing well when compared over RNN model. In 2020, Maqsood *et al*. [16] have explored the impact of various events happened in the year 2012-2016 on stock markets. Here, Twitter dataset was employed for computing the sentiment analysis to each of these events. The dataset included millions of tweets,which were employed for defining the event sentiment.In 2019, Abdi *et al*. [17] have suggested a deep-learning-based technique for categorizing the opinion of the user mentioned in reviews. Moreover, a deep learning model was a unified feature set that was representative of sentiment shifter rules, word embedding, sentiment knowledge, linguistic and statistical knowledge has not been continuously explored for a sentiment analysis. Moreover, the suggested model used RNN that consisted of LSTM for considering the benefit of sequentialprocessing and conquered many issues in conventional algorithms. In 2020, Park *et al*. [18] havedesigned a deep learning approach for improving performance. In order to improve the performance, two questions have come into picture. The content attention was required for being sophisticated formerging many attention results non-linearly and assumes the whole context for mentioning thecomplex sentences. The test results have shown that the proposed model was attained as the bestperformance.

3.Objective of Research:

- We will develop sentiment analysis algorithm that overcome the current problem like insufficient labeled data and incapability to deal with complex sentences.
- We will work on sarcasm detection, double negative detection, and multipolarity. To train and test proposed model ,web use biggest open-source database for sentiment analysis created by IMDb.
- Through sentiment analysis of movie reviews, customer feedback, support tickets, we will be able to explore the different characteristics of product, which indirectly will help companies to discover many interesting things.These steps sentiment analysis in our research:

1. Collect feedback data(Data collection)
2. Make sure data is qualitative enough for analysis.
3. Data processing.
4. Data analysis.
5. Data visualization.

4.Methodology to be adopted:

Data collection and preparation. First, you need to gather all relevant brand mentions in one document. Consider selection criteria – should these mentions be time-limited, use only one language, come from a specific location, etc. Then data must be prepared for analysis: one has to read it, delete all non-textual content, fix grammar mistakes or typos, exclude all irrelevant content like information about reviewers, etc. Once we have data prepared, we can analyze it and extract sentiment from it. You can learn more about data preparation in our story.

As dozens or even hundreds of thousands of mentions may require analysis, the best practice is to automate this tedious work with software.

Using off-the-shelf tools and APIs. Various customer experience software (e.g. InMoment, Clarabridge) collect feedback from numerous sources, alert on mentions in real-time, analyze text, and visualize results. Text analysis platforms (e.g. DiscoverText, IBM Watson Natural Language Understanding, Google Cloud Natural Language, or Microsoft Text Analytics API) have sentiment analysis in their feature set.

InMoment provides five products that together make a customer experience optimization platform. One of them, Voice of a Customer, allows businesses to collect and analyze customer feedback in a text, video, and voice forms. The number of data sources is sufficient and includes surveys, social media, CRM, etc. Developers provide users with real-time notifications, custom dashboards, and various reporting options.

Clarabridge is a customer experience management (CEM) platform. It pulls and analyzes text from chats, survey platforms, blogs, forums, and review sites. Users can also gain insights from emails, employee and agent notes, call recordings and Interactive Voice Response (IVR) surveys: The system can convert them into text. They provide social media listening as well. The system considers industry and source, understanding the meaning and context of every comment. Sentiment analysis results display on an 11-point scale. Users can modify sentiment scores to be more business-specific if needed.

DiscoverText is a cloud-based collaborative text analytics system for researchers, entrepreneurs, and governments. Capterra users note the solution is great for importing/retrieving, filtering, and analyzing data from various sources, including Twitter, SurveyMonkey, emails, and spreadsheets. Sentiment analysis is one of numerous text analysis techniques of DiscoverText.

IBM Watson Natural Language Understanding is a set of advanced text analytics systems. Analyzing text with this service, users can extract such metadata as concepts, entities, keywords, as well as categories and relationships. It also allows for defining industry and domain to which a text belongs, semantic roles of sentence parts, a writer's emotions and sentiment change along the document. IBM Watson Natural Language Understanding currently supports analysis in 13 languages. Tools for developers are also provided, so they can build their solutions (e.g. chatbots) using IBM Watson services.

5.Importance of society applications:

Use cases of sentiment analysis

Various industries use sentiment analysis. While the areas of sentiment analysis application are interconnected, they are all about enhancing performance via analysis of shifts in public opinion.

- Brand monitoring

If the Internet was a mountain river, then analyzing user-generated content on social media and other platforms is like fishing during trout-spawning season. People enjoy sharing their points of view regarding the latest news, local and global events, and their experience as customers. Twitter and Facebook are favorite places for daily comment wars and spirited (to put it mildly!) conversations. News about celebrities, entrepreneurs, and global companies draw thousands of users within a couple of hours after being published on Reddit. Media giants like *Time*, *The Economist*, CNBC, as well as millions of blogs, forums, and review platforms flourish with content on various topics.Why not use these data sources to monitor what people think and say about your organization and why they perceive you this way? Sentiment analysis of brand mentions allows you to keep current with your credibility within the industry, identify emerging or potential reputational crises, to quickly respond to them. You can compare this month's results and those from the previous quarter, for instance, and find out how your brand image has changed during this time.It's not only important to know social opinion about your organization, but also to define who is talking about you. Measuring mention tone can also help define whether industry influencers are mention your brand and in what context. And what's more exciting, sentiment analysis software does all of the above in real time and across all channels.

- Competitive research

There is one thing for sure you and your competitors have in common – a target audience. You can track and research how society evaluates competitors just as you analyze their attitude towards your business. What do customers value most about other industry players? Is there anything competitors lack or do wrong? Which channels do clients use to engage with other companies? Use this knowledge to improve your communication and marketing strategies, overall service, and provide services and products customers would appreciate.Competitive analysis that involves sentiment analysis can also help you understand your weaknesses and strengths and maybe find ways to stand out.

- Flame detection and customer service prioritization

Hospitality brands, financial institutions, retailers, transportation companies, and other businesses use sentiment classification to optimize customer care department work. With text analysis platforms like IBM Watson Natural Language Understanding or Monkey learn, users can automate the classification of incoming customer support messages by polarity, topic, aspect, and priority. Then queries are sent to dedicated teams and specialists. Since it's better to put out a spark before it turns into a flame, new messages from the least happy and most angrycustomers are processed first , for example, groups feedback by device, customer journey stage, and new or repeat customers.

- Product analysis

Every entrepreneur dies to see fans standing in lines waiting for stores to open, so they can run inside, grab that new product, and become one of the first proud owners in the world. How to bring the desired product to market? The only approach is to ask people what they want. Successful companies build a minimum viable product (MVP), gather early feedback, continuously improving a product even after its release. Feedback data comes from surveys, social media, and forums, and interaction with customer support. Questions like how to define which customer groups to ask, analyze this ocean of data, and classify reviews arise.That's when sentiment analysis comes in handy. It allows for learning about product advantages and drawbacks. For example, a student from Oklahoma State University has analyzedAmazon reviews about two Samsung phone models (Galaxy S6 & Galaxy S7) and two devices from Apple (iPhone 6 & iPhone 7) to find out why customers preferred one brand over another. He found out that users whose priorities are a reliable battery and a good screen choose Samsung phones. And customers who are more interested design and camera buy iPhones. Filtering comments by topic and sentiment, you can also find out which features are necessary and which must be eliminated. Armed with sentiment analysis results, a product development team will know exactly how to deliver a product that customer would buy and enjoy.

- Market research and insights into industry trends

As we said before, social media sites and forums are sources of information on any topic. People discuss news and products; write about their values, dreams, everyday needs, and events. And they do this voluntarily 24/7.Sentiment analysis solves the problem of processing large volumes of unstructured data. Using this type of text analysis, marketers track and study consumer behavior patterns in real time to predict future trends and help management make informed decisions. Another benefit of sentiment analysis is that it doesn't require heavy investment and allows for gathering reliable and valid data since its user-generated.

- Workforce analytics/employee engagement monitoring

Some organizations go beyond using sentiment analysis for market research or customer experience evaluation, applying it internally for HR-related processes. These companies measure employee satisfaction, detect factors that discourage team members and eventually reduce company performance. Specialists automate the analysis of employee surveys with SA software, which allows them to address problems and concerns faster. Human resource managers can detect and track the general tone of responses, group results by departments and keywords, and check whether employee sentiment has changed over time or not.Sentiment analysis takes employee mood monitoring to the next level with real-time monitoring capabilities. For instance, team members can fill out survey forms with a single request to rate their workplace conditions every month. They can also analyze their posts in social media to find a possible connection between their state of mind and work lives.

6.Proposed work plan:

The main problems that exist in the current techniques are: inability to perform well in different domains, inadequate accuracy and performance in sentiment analysis based on insufficient labeled data, incapability to deal with complex sentences that require more than sentiment words and simple analyzing.

We will develop sentiment analysis algorithm that overcome the current problem like insufficient labeled data and incapability to deal with complex sentences.

These steps sentiment analysis in our research:

1. Collect feedback data
2. Make sure data is qualitative enough for analysis

Sentiment analysis allows businesses to harness tremendous amounts of free data to understand customer needs and attitude towards their brand. Organizations monitor online conversations to improve products and services and maintain their reputation. The analysis takes customer care to the next level. Customer support systems with incorporated SA classify incoming queries by urgency, allowing employees to help the most demanding customers first. Sentiment analysis is a powerful tool for workforce analytics as well.

The approximate duration proposed to complete research project in 36 months. Presentation will be done after every 6 months

- Literature review-after 6 months and Implementation of basic preprocessing algorithms.
- Implementation of available algorithms and proposed methods.
- Comparing the results obtained from proposed and existing methods.

Final submission of thesis will be after permission of guide and fulfilling minimum requirements.

7.References

[1] B. Gokulakrishnan, P. Priyanthan, T. Ragavan, N. Prasath and A. Perera, "Opinion mining and sentiment analysis on a Twitter data stream", *IEEE*, 2012.
[2] I. Maks and P. Vossen, "A lexicon model for deep sentiment analysis and opinion mining applications", *ScienceDirect*, 2012.
[3] Wehrmann Joonatas, Willian Becker, Henry EL Cagnini and Rodrigo C. Barros, "A character-based convolutional neural network for language-agnostic Twitter Sentiment Analysis", *IEEE*, 2017.
[4]Duric and F. Song, "Feature selection for sentiment analysis based on content and

syntax models", *ScienceDirect*, 2012 .
[5]Cambria Erik et al., "New avenues in opinion mining and sentiment analysis", *IEEE Intelligent systems*, vol. 28, no. 02, pp. 15-21, Feb 2013.
[6] Sneha A. Sahare, "A Survey Work on Opinion Mining and Sentiment Analysis", *International Journal of Advance Research and Innovative Ideas in Education (IJARIIE)*, vol. 3, no. 2, 2017.
[7] S. ChandraKala and C. Sindhu, "Opinion Mining And Sentiment Classification: A Survey".
[8] S. E. Saad and J. Yang, "Twitter Sentiment Analysis Based on Ordinal Regression," IEEE Access, vol. 7, pp. 163677-163685, 2019.
[9] Y. Fang, H. Tan and J. Zhang, "Multi-Strategy Sentiment Analysis of Consumer Reviews Based on Semantic Fuzziness," IEEE Access, vol. 6, pp. 20625-20631, 2018.
[10] A. Feizollah, S. Ainin, N. B. Anuar, N. A. B. Abdullah and M. Hazim, "Halal Products on
Twitter: Data Extraction and Sentiment Analysis Using Stack of Deep Learning Algorithms," IEEE Access, vol. 7, pp. 83354-83362, 2019.
[11] ParamitaRay, and AmlanChakrabarti, "A Mixed approach of Deep Learning method and Rule-Based method to improve Aspect Level Sentiment Analysis", Applied Computing and Informatics, Available online 4 March 2019.
[12] Mohammad A.Hassonah, RizikAl-Sayyed, AliRodan, Ala' M.Al-Zoubi, IbrahimAljarah, andHossamFaris, "An efficient hybrid filter and evolutionary wrapper approach for sentiment analysis of various topics on Twitter", Knowledge-Based Systems, vol. 192, 15 March 2020.
.

Swati Bapurao Tondare
Research Scholar

Dr. Archana Harsing Sable
Research Supervisor